Modeling Subject Scoring Behaviors in Subjective Experiments Based on a Discrete Quality Scale

Lohic Fotio Tiotsop, Antonio Servetti, Marcus Barkowsky, Enrico Masala

VQEG JEG-Hybrid

Jul 2024

Introduction

- Raw ratings from subjects are typically noisy
  - Subject fatigue or distraction
  - Complex stimuli can impact the accuracy of naive raters
  - Presence of spammer annotators
- Statistical models for subjective quality recovery and peculiar behavior identification
  - Different approaches have been proposed
  - Subjects are commonly assumed to exhibit bias and inconsistency
- Our work adopts this common perspective, but:
  - Rather than an overall bias, we define **positional** bias weights
  - Subject inconsistency arises from a **scoring model** that is derived, **not assumed a priori**

Notation

- $\mathcal{I}$: the set of stimuli that have been rated;
- $\mathcal{J}$: the set of subjects that rated the stimuli in $\mathcal{I}$;
- $\mathcal{K}$: the set of opinion scores available on the quality scale;
- $\mathcal{F}$: the set of influence factors that might affect the ratings of a subject;
- r_i^j: the rating of the subject $j \in \mathcal{J}$ for the stimulus $i \in \mathcal{I}$;
- $\mathcal{R}$: all the ratings collected during the subjective test;
- n_{ik}: the number of subjects in $\mathcal{J}$ that chose the opinion score $k \in \mathcal{K}$ for the stimulus $i \in \mathcal{I}$.

Subjective Quality Recovery

- The MOS of stimulus $i \in \mathcal{I}$ is:

$$MOS_i = \sum_{j \in \mathcal{J}} \frac{1}{|\mathcal{J}|} \cdot r_i^j = \sum_{k \in \mathcal{K}} \frac{n_{ik}}{|\mathcal{J}|} \cdot k \qquad (1)$$

- The MOS weights the opinion score k with $\frac{n_{ik}}{|\mathcal{J}|}$
- This weighting schema is not robust to noisy ratings
- We define the ground-truth quality of stimulus i as:

$$Q_i = \sum_{k \in \mathcal{K}} w_{ik} \cdot k, \qquad (2)$$

where the weights w_{ik} are to be computed, taking into account the noisy nature of the gathered data.

Regularized Maximum Likelihood Estimation (RMLE) of Quality

- The weight w_{ik} can be assimilated to the actual probability of scoring stimulus i with k, thus the Log likelihood function is: $LL(w) = \sum_{i\in\mathcal{I}} \sum_{k\in\mathcal{K}} n_{ik} \cdot \log(w_{ik})$
- The classical MLE approach would yield the not robust solution $w_{ki} = \frac{n_{ik}}{|\mathcal{J}|}$
- We added a regularization term to the likelihood function to account for noise
- **The regularization term penalizes not frequently chosen opinion scores on the scale**:

$$R(w) = \sum_{i\in\mathcal{I}} \sum_{k\in\mathcal{K}} C_{ik} \cdot w_{ik}, \quad (3)$$

where $C_{ik} = -log\left(\frac{n_{ik}}{|\mathcal{J}|}\right)$

Regularized Maximum Likelihood Estimation (RMLE) of Quality

Definition

The weights w_{ki} yielding the RMLE estimation of the quality the stimuli in $\mathcal{I}$ on the discrete quality scale $\mathcal{K}$ are the optimal solution of the following problem:

$$\begin{aligned} &\max_{w} \left[LL(w) - \lambda \cdot R(w) \right] \\ &s.t. \quad \sum_{k \in \mathcal{K}} w_{ik} = 1 \quad \forall i \in \mathcal{I} \end{aligned} \tag{4}$$

Where $\lambda = \frac{1}{2} \cdot \frac{|\mathcal{I}||\mathcal{K}|}{|\mathcal{J}|}$ is a regularization coefficient

Positional Bias Weights

- A single overall bias might not be enough to highlight certain peculiar behavior
- The following behaviors might be observed in subjective test on a discrete scale:
  1. **Positively biased annotators**;
  2. **Negatively biased annotators**;
  3. **Unary annotators**;
  4. **Binary annotators**;
  5. **Ternary annotators**;
  6. **Adversary annotators**;
  7. **Spammer annotators**;
  8. **Competent annotators**.
- 3, 4 and 5 suggest that a subject might prefer one or certain opinion scores more than others

Positional Bias Weights

- We introduce μ_k^j as the systematic tendency of subject $j \in \mathcal{J}$ to prefer the opinion score k over the others
- We performed one-hot encoding of subject ratings to estimate the value of μ_k^j:

$$R_i^j(k) = \begin{cases} 1 & \text{if } k = r_i^j \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

- μ_k^j is estimated as:

$$\mu_k^j = \frac{\sum_{i \in \mathcal{I}} \left(R_i^j(k) - w_{ik} \right)}{|\mathcal{I}|}. \tag{6}$$

Positional Bias Weights

- Note that it holds:

$$\sum_{k \in \mathcal{K}} \mu_k^j = 0 \quad \forall j \in \mathcal{J} \tag{7}$$

- Thus, some bias weights of a subject j are positive and others are negative
  - if $\mu_k^j > 0$, then subject j tends to prefer k as opinion score
  - if $\mu_k^j < 0$, then subject j tends not to select k as opinion score
- The overall bias of subject $j \in \mathcal{J}$ can also be estimated as:

$$b_j = \sum_{k \in \mathcal{K}} k \cdot \mu_k^j. \tag{8}$$

Deriving the Scoring Model & the Subject Inconsistency

- Previous approaches assume a priori a probabilistic scoring model
- Here, higher-level assumptions are made and a scoring model is formally derived
- Our idea: **subjects unconsciously attribute a stochastic attractiveness to each opinion score on the quality scale and choose the one with the highest perceived attractiveness**
- The attractiveness of each opinion score depends on:
  - The stimulus actual quality
  - The subject tendency to select that opinion score
  - Numerous stochastic and thus uncontrollable influence factors

Definition of Attractiveness

Definition

The attractiveness of the opinion score k for the subject j when rating the stimulus i is defined as:

$$U^j_{ik} = w_{ik} + \mu^j_k + \theta^j_{ik}, \tag{9}$$

where θ^j_{ik} is a random variable modeling the relevance of the effect of all the influence factors.

- In practice the distribution of θ^j_{ik} is unknown
- Some mild assumptions on it are required to derive our scoring model

Modeling the Effect of Influence Factors (IF)

- Let us denote by θ^j_{ikf} the relevance of the effect of the IF $f \in \mathcal{F}$
- We assume that the subject is mainly influence by the IF with the largest relevance, thus $\theta^j_{ik} = \max_{f \in \mathcal{F}} \theta^j_{ikf}$
- We further assume that the distribution of each random variable θ^j_{ikf} has a heavy tail. Denoting by $F^j_{ik}(x)$ the unknown cumulative probability distribution of any random variable θ^j_{ikf} $f \in \mathcal{F}$. We assume there exist two constants $\alpha_{|\mathcal{F}|}$ and $\beta_j > 0$ such that, $\forall i \in \mathcal{I}, \forall j \in \mathcal{J}, \forall k \in \mathcal{K}$:

$$\lim_{|\mathcal{F}| \to +\infty} F^j_{ik} \left(\frac{1}{\beta_j} x + \alpha_{|\mathcal{F}|} \right)^{|\mathcal{F}|} = \exp\left(-e^{-x}\right) \quad \forall x \in \mathbb{R}. \quad (10)$$

- β_j is related to the probability distribution of IFs and thus to the inconsistency of subject j
- These assumptions do not really limit the model's application scope

Deriving the Scoring Model

- In practice, the number of IFs is very large
- The following Theorem yields our scoring model:

Theorem

As the number of IFs tends to infinity, i.e., $|\mathcal{F}| \to +\infty$, the probability that subject j chooses opinion score k when rating stimulus i is:

$$p_{ik}^{j} = \frac{e^{\beta_j (w_{ik} + \mu_k^j)}}{\sum_{k \in \mathcal{K}} e^{\beta_j (w_{ik} + \mu_k^j)}}, \quad k \in \mathcal{K}, \quad j \in \mathcal{J}, \quad i \in \mathcal{I}. \qquad (11)$$

- Thus r_i^j is a $|\mathcal{K}|$-class discrete random variable and theorem provides its density

Link Between β_j & the Subject Inconsistency

- The closer to 0 β_j is, the more inconsistent is subject j
- But β_j alone might not fully capture all aspects of subject j inconsistency
- The inconsistency σ_{ij}^2 of subject j on the quality of stimulus i is defined as the variance of r_i^j:

$$\sigma_{ij}^2(\beta, \mu, w) = \sum_{k \in \mathcal{K}} k^2 \cdot p_{ik}^j - \left(\sum_{k \in \mathcal{K}} k \cdot p_{ik}^j \right)^2 \qquad (12)$$

- The overall inconsistency of subject j is then:

$$\sigma_j^2(\beta, \mu, w) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \sigma_{ij}^2(\beta, \mu, w) \qquad (13)$$

β_j Estimation

- β_j is estimated by performing a least square fitting of the model's variance to the observed variance of the ratings of subject j
- The observed variance is computed as:

$$s_j^2 = \mathrm{Var}(Q - R^j) \tag{14}$$

  where R^j represents all the rating given by the subject j and Q the recovered qualities of the stimuli.
- β_j is estimated as the value that minimizes the function $l(\beta_j)$ defined as:

$$l(\beta_j) = \left(s_j^2 - \sigma_j^2(\beta_j, \mu, w)\right)^2 \tag{15}$$

Results

- The results of the typical experiments on the robustness to noise of quality recovery approaches are omitted here, (please refer to the paper if interested).

Journals & Magazines > IEEE Transactions on Multimedia > Early Access

Modeling Subject Scoring Behaviors in Subjective Experiments Based on a Discrete Quality Scale

Publisher: IEEE Cite This PDF

Lohic Fotio Tiotsop ; Antonio Servetti ; Marcus Barkowsky ; Enrico Masala   All Authors

33
Full
Text Views

Open Access

Under a Creative Commons License

Abstract
Authors
Keywords
Metrics

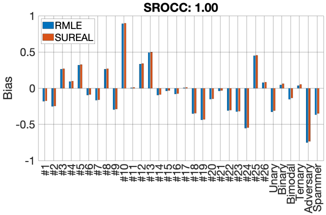
Abstract:
Several approaches have been proposed to estimate quality in subjective experiments while highlighting peculiar subject behaviors. However, there is some room for improvement in existing approaches, both in terms of robustness to noise and the ability to accurately indicate several peculiar subject behaviors in subjective experiments. This work advances the state-of-the-art in three main directions: i) A new approach to estimate the subjective quality from noisy ratings is proposed and is shown to be more robust to noise than are four state-of-the-art approaches; ii) a novel subject scoring model is proposed that makes it possible to highlight several peculiar behaviors typically observed in subjective experiments; and iii) our proposed probabilistic subject scoring model results from the proof of a theorem, whereas in previous approaches a probabilistic scoring model is assumed a priori . This represents an important first step toward models supported by a stronger theoretical foundation. Numerical experiments conducted on several datasets highlight the effectiveness of our proposal.

Published in: IEEE Transactions on Multimedia (Early Access)

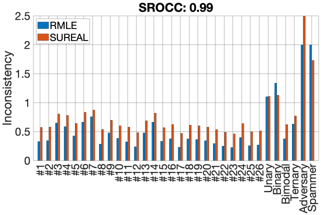
- Instead, let us analyse the potential of the scoring model to highlight peculiar behaviors on discrete scales.

Peculiar Behaviors Identification: Sureal vs Proposal

- Experiments done on the Netflix Public datasets integrated with the simulation of six peculiar behaviors
- SUREAL and the proposed approach are well aligned in terms of overall bias and inconsistency



(a) Subject overall bias (b_j)



(b) Subject overall inconsistency ($\sigma_j^2(\beta, \mu, w)$)

- However, the proposed approach brings new insights into the explanation of the source of the observed peculiarities

Bias Weights Analysis

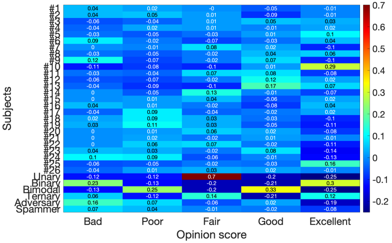


Figure: Subject bias weights (μ_k^j) computed on the Netflix Public dataset integrated with six simulated peculiar subjects

- Subject #10 favors the higher end of the quality scale
- Subject #6 prefers the quality scale extremes
- Subject #7 seldom chooses "excellent" without compensating by selecting "good"
- Subject #14 seems a unary annotator
- The proposed approach can perfectly highlight simulated subjects with positional bias

Analysis of the Subject's Inconsistency

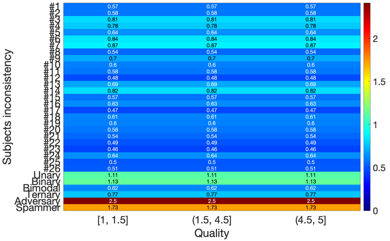


Figure: SUREAL subject inconsistency as function of the recovered quality computed on the Netflix Public dataset integrated with six simulated peculiar subjects

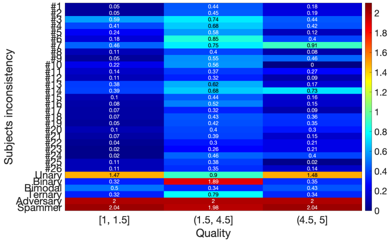


Figure: Proposed model subject inconsistency ($\sigma_j^2(\beta, \mu, w)$) as function of the recovered quality computed on the Netflix Public dataset integrated with six simulated peculiar subjects

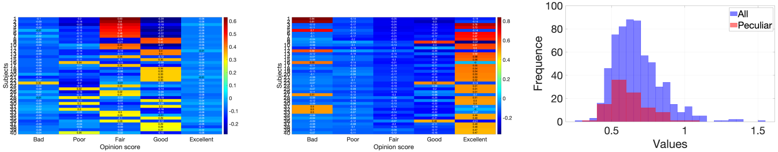
- Modeling higher accuracy of subjects at the quality scale extremes
- Automatically highlighting where a subject is inconsistent

Analysis of Two Crowdsourcing Datasets

1. KoNViD-1k VQA Database: opinion scores of 624 subjects on a five point scale regarding the **perceptual quality** of 1200 short video sequences
2. MovieLens 1M Dataset: opinion scores of 6040 subjects on a five point scale regarding their **overall satisfaction** with 3952 movies.

- We analysed the frequency of occurrence of some peculiar behaviors in crowdsourcing tests due to positional bias.
- We evaluated whether the subject's inconsistency alone can highlight these peculiar behaviors.

Unary Behavior: Tendency to Prefer one Opinion Score



(a) KoNVID: top 40 unary annotators

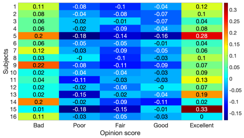
(b) MovieLens: top 40 unary annotators

(c) KoNVID: Sureal subject inconsistency

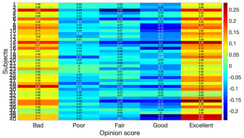
Figure: Unary behavior analysis

- Unary behavior/pattern seems to be quite frequent: around 16% of subjects on average;
- Unary annotators seem to be hardly identifiable through the subjects' inconsistency.

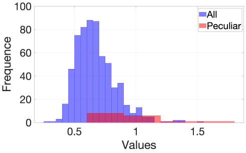
Binary Behavior: Tendency to Prefer Extreme Opinion Scores



(a) KoNVID: all binary annotators



(b) MovieLens: top 40 binary annotators

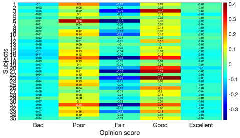


(c) KoNVID: Sureal subject inconsistency

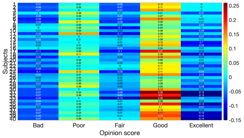
Figure: Binary behavior analysis

- Binary behavior is not very frequent: around 2% on average;
- Bias weights highlights the source of the high inconsistency of binary annotators

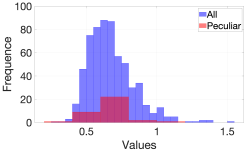
Bimodal Behavior: Scoring with Mode around 'Poor' and 'Good'



(a) KoNVID: all bimodal annotators



(b) MovieLens: top 40 bimodal annotators



(c) Sureal subject inconsistency

Figure: Bimodal behavior analysis

- bimodal behavior frequency: around 4% on average;
- Subject inconsistency do not highlight the peculiarity induced by a bimodal scoring behavior.

Conclusions

Results summary

- The positional bias weights enable a more comprehensive analysis of subjects behavior
- The derived scoring model capture the higher accuracy of subjects at the extremes of the quality scale
- The proposed model can potentially enable an efficient visual analysis of large-scale datasets.

Open questions

- Finding a numerically stable approach to fit the model to data and thus estimate both stimuli quality and subjects characteristics at once
- Any other interesting future directions?

Thanks for your attention